DOCUMENT RESUME

ED 210 283 TM 810 884

TITLE Speaking Assessment, 1981. Final Report.
INSTITUTION Massachusetts State Dept. of Education, Boston. Bureau of Research and Assessment.
PUB DATE Aug 81
NOTE 85p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Achievement Rating; Grade 8; *Inservice Teacher Education; Junior High Schools; *Observation; *Speech Communication; Speech Evaluation; Speech Skills; *Student Evaluation; Training Methods
IDENTIFIERS *Basic Skills Improvement Policy (Massachusetts)

ABSTRACT

The Bureau of Research and Assessment developed a two-staged approach for assessing speaking skills. All students were to be initially rated by two of their current classroom teachers-(teacher observations). A limited number of students would be assessed by a rater in a focused measure of student performance (one-on-one ratings). The purpose of the 1981 study was improvement of materials used to train teachers and individual raters. Training materials were substantially revised. Reliability of teacher ratings and individual raters were studied. The one-on-one rating was found to be a more valid measure of speaking skills. It is the preferred method because it is standardized and a reliable measure when administered by trained, approved raters. Results can be reviewed because the measure is obtained in a controlled setting. If necessary, ratings can be conducted by an outside agency and can be audited and verified. The one-on-one rating measure is not ready for dissemination at this time. Training materials need more revision and the method for determining whether raters will be reliable needs to be expanded and tested. (DWH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. DiBella

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Massachusetts Board of Education

Anne H. McHugh, Chairperson
James L. Greca, Vice Chairperson

John W. Bond
Millie Clements
James P. Doherty
James R. Grande
Howard A. Greis
Charles T. Grigsby
Mary Ann Hardenbergh
Armando Martinez
Edwin M. Rossman
Donald R. Walker

Gregory R. Anrig, Commissioner of Education, Secretary

Bureau of Research and Assessment

Allan S. Hartman, Director
Tracy Libros
Leslie S. May
Matthew H. Towle
Betty Hancock

The Massachusetts Department of Education insures equal employment/educational opportunities/and affirmative action, regardless of race, color, creed, national origin, or sex, in compliance with Title IX, or handicap, in compliance with Section 504.

## EXECUTIVE SUMMARY

### THE STUDY

Entering this year's study, the Bureau of Research and Assessment was envisioning a two-staged approach to speaking assessment. The tentative plan called for students to be initially rated by two of their current classroom teachers ("teacher observations") and then a limited number, to be chosen either because of the close proximity of their scores to the cut-score or because of substantial disagreement between the teachers, would be assessed by rater in a focused measure of student performance ("one-on-one ratings"). The purpose of this year's study was to improve the materials used to train the teachers and one-on-one raters and to determine the decision rule for requiring one-on-one rating that would minimize the number of individual assessments without seriously affecting reliability.

In order to carry out this purpose, the training materials were substantially revised and then used to train teachers and one-on-one raters in eight schools throughout Massachusetts. In each school, sixty students were selected. Each student received ratings from four teacher observers and two one-on-one raters. This way, all data needs were duplicated, and it was possible to investigate the reliability issues in some detail.

### RESULTS

The results of the study showed:

- The one-on-one ratings were satisfactorily reliable in half the schools.
- The results of the practical exercise given at the end of the training session predicted with substantial accuracy which pairs of raters would be reliable.
- The teacher observation data exhibited poor reliability in most schools.
- There appeared to be nothing in the background information collected which could predict those few cases where the teacher observation showed sufficient reliability.
- The one-on-one measure and teacher observation measure are two entirely different measures, with a low intercorrelation.

## CONCLUSIONS

The results indicate that it makes no sense to use the teacher observation and one-on-one rating measures in tandem. Consequently, one measure must be selected for use. The one-on-one rating system is preferable, for the following reasons:

1. It is a more valid measure of speaking skills.
2. It is standardized.
3. Because the measure is obtained in a controlled setting, the results can be reviewed.
4. It is a reliable measure when administered by trained, approved raters.
5. It can be conducted by an outside agency, if necessary, and results can be audited and verified.

The one-on-one rating measure is not yet ready for final dissemination. The training materials need more revision, the method for determining whether raters will be reliable needs to be expanded and tested, and it must be determined if the process can be conducted without racial/ethnic bias. A study to accomplish these tasks should be conducted next year.

TABLE OF CONTENTS


| | | Page |
|---|---|---|
| | EXECUTIVE SUMMARY | i |
| I. | BACKGROUND | 1 |
| II. | DEVELOPMENT OF MATERIALS | 3 |
| III. | TEACHER OBSERVATION TRAINING | 6 |
| IV. | ONE-ON-ONE TRAINING | 8 |
| V. | DATA COLLECTION FOR THE TEACHER OBSERVATIONS AND ONE-ON-ONE ASSESSMENTS | 10 |
| VI. | ANALYSIS OF WORKSHOP DATA | 11 |
| VII. | ANALYSIS OF THE FIELD DATA | 16 |
| VIII. | COMPARING THE TEACHER OBSERVATION MEASURE TO THE ONE-ON-ONE MEASURE | 19 |
| IX. | CHOOSING BETWEEN THE TWO MEASURES | 23 |
| X. | RECOMMENDATIONS | 25 |
| | APPENDICES A-F | 27 |

TABLE OF TABLES


| | | Page |
|---|---|---|
| Table 1 | Ratings of Student 1 Made by 97 Workshop Teachers, Reported by Dimension | 11 |
| Table 2 | Number of Times Teacher Ratings Agreed with Expert Ratings, Reported by Student | 12 |
| Table 3 | Number of Times Teacher Ratings Agreed with Expert Ratings, Reported by Dimension | 12 |
| Table 4 | Total Scores Assigned by the 97 Workshop Teachers to Each of the Five Sample Students | 13 |
| Table 5 | The Number of Times One Sample Student Was Scored Higher Than Another by the 97 Workshop Teachers | 14 |
| Table 6 | Descriptive Statistics of the Consistency of Judgments Made by Pairs of One-on-One Raters | 17 |
| Table 7 | Correlations Between Observation Ratings by Teachers in Various Subject Matter Areas and One-on-One Ratings | 21 |
| Table 8 | The Reliability of the One-on-One Raters, Using a Total Score on All Four Tasks as Measure, and Using a Total Score on Tasks 1 and 4 as a Measure, Reported by School | 24 |

# I. BACKGROUND

## 1979-1980

As a part of the Basic Skills Improvement Policy, the Bureau of Research and Assessment developed a two-staged approach for assessing speaking skills. This approach included:

- Teacher observation ratings

- One-on-one ratings

The teacher observation ratings are a general measure of a student's speaking performance. In this approach, two teachers who have a student currently enrolled in class independently rate the student's general speaking performance using the speaking rating scales. Teachers base their ratings of a student on observations of the student's performance in normal classroom activities, such as asking questions, responding to questions, explaining how to do something, giving a report to the class or talking with other students in discussion groups. The teachers consider the student's performance on the average since the beginning of the current semester.

The one-on-one ratings are a focused measure of a student's speaking performance. In this approach, one specially trained rater rates the student's performance using similar speaking rating scales to those used in the teacher observation approach. However, instead of basing ratings on classroom observations, the one-on-one rater gives the student four specific speaking tasks and rates the student's oral responses to these tasks.

The original rationale for developing two approaches for assessing speaking was based on the need for measures that are feasible for large-scale use by school districts in the future. The intention was to use the teacher observation approach as a screening measure and to use the one-on-one approach as a back-up measure for students whose level of ability was in question.

The two-staged approach was field tested during the spring of 1980 with a random sample of 12th graders from 49 schools across the state. The results of the study were promising, and provided a statewide overview of students' speaking skills. However, the results indicated a need for some changes in the assessment instruments, particularly a need for increased variability in both the teacher observation and the one-on-one ratings; a need for detailed training materials so that the assessments could be implemented easily in schools in the future; and also, a need to study the issue of reliability of these instruments in more detail.

1980-81

The purpose of the 1980/81 study was to test the feasibility and reliability of the two-staged assessment when implemented in settings similar to those in which assessments would be implemented in the future. The tasks of the study were to refine the teacher observation and one-on-one instruments, to develop appropriate training materials and to conduct trial assessments of eighth graders in several representative schools. Eighth graders were selected because many districts are conducting basic skills assessment at this grade.

In order to test the reliability of the assessment, a replication study was designed. In this replication study all of the assessment ratings were duplicated. Instead of conducting two teacher observations for each student, four teacher observations were obtained; instead of conducting a single one-on-one rating, two one-on-one ratings were obtained. The replicated information was used to measure the reliability of the instruments. The observation ratings of two of the teachers were compared to ratings of the other two teachers to see if both pairs of teachers rated the student in the same way. Similarly, the one-on-one ratings of one rater were compared to the one-on-one ratings of the other rater to see if both raters rated the student in the same way.

The project was conducted in eight schools which represented a variety of school settings. Three schools were located in large cities with minority student populations, three in suburban areas and the remaining two in primarily rural areas. Within each school 60 eighth grade students were randomly selected to participate in the project. Each selected student received observation ratings from four of his or her current teachers. In addition, each student received one-on-one ratings from two trained raters from the school district.

## II. DEVELOPMENT OF MATERIALS

The major effort of the development and revision of materials was aimed toward increasing the reliability of both the teacher observation and one-on-one approaches.

### SPEAKING RATINGS GUIDE

The first step of this effort was the refinement of the Speaking Ratings Guide. The guide was divided into separate components: a Teacher Observation Ratings Guide and a One-on-One Ratings Supplement. The two versions of the guide remain parallel but describe the levels of performance in slightly different ways (the new versions are provided as Appendix A). The Teacher Observation Ratings Guide focuses on students' speaking performance in normal classroom interactions over the course of the semester. The One-on-One Ratings Supplement focuses on students' speaking performance in response to each of the four speaking tasks.

Both versions of the guide were expanded to a five-point scale. The five points of the scale take into account the level (poor, satisfactory, or good) and the consistency (most of the time or some of the time) of the student's response. A complete explanation is included in the guide (see Appendix A).

The four dimensions -- delivery, organization, content, language -- which represent the major components of the speaking process, remained unaltered in the guides. However, the descriptions within each of the levels of the dimensions were refined so that distinctions might be made more readily.

### AUDIOTAPES

In addition to the written guides, audiotapes were developed to exemplify the levels of the dimensions. Specifically, 22 taped examples of students spontaneous responses to typical classroom speaking tasks were produced to accompany the Teacher Observation Ratings Guide. Twenty-four taped examples were produced for the One-on-One Ratings Supplement. These examples were from spontaneous responses to the four tasks used in the 1980 statewide assessment. The tapes provided concrete examples of the poor, satisfactory and god levels for each of the four dimensions.

An additional 45 taped examples were produced to allow teachers and one-on-one test administrators to practice rating students' responses. Twenty-five of these examples were used to conduct inter-rater reliability checks. Many of the examples were selected from the existing tape pool acquired during the 1979/80 statewide assessment. Other tapes, particularly

those developed as examples for the Teacher Observation Ratings, were chosen from students' responses in typical classroom activities. Responses that exhibit various forms of speech and dialect of students from a variety of racial/ethnic backgrounds were selected.

## OTHER MATERIALS

Training scripts, with accompanying overhead transparencies, were developed for both the Teacher Observation Training sessions and the One-on-One Training session. A background questionnaire for teachers and one-on-one raters was also developed and distributed at these sessions. The questionnaire requested the following information: 1) sex; 2) primary teaching subject; and 3) past course work in teaching oral communication skills. A copy of the questionnaire is provided in Appendix B.

A Teacher Observation Ratings form was designed for the teacher observation assessment. This form consists of a 1-5 scale for each of the four dimensions. For experimental purposes, teachers were asked to provide an overall rating of the student's speaking skills on a scale of one to eight. The two rating approaches were used to determine which method yields higher reliability. Teachers were also asked to indicate how familiar they were with the speaking skills of the student, how often the student speaks in class and whether there were special circumstances that made the teacher unable to rate a student's speaking skills. These questions were included to determine the degree of familiarity and confidence teachers have in rating students speaking skills. A copy of this rating form is provided in Appendix C.

Additional materials developed for the one-on-one assessment included a Test Administrator s Manual and One-on-One Rating forms. The manual consists of step-by-step instructions to guide the rater through a testing session. It includes a checklist of necessary materials, requirements for an adequate testing location, the four speaking tasks and specific instructions to be given to students in order to ensure standardized test administrations. Raters entered their ratings of 1-5 for each of the dimensions for each task on the ratings sheet. Copies of these materials are found in Appendices D and E.

A background questionnaire was designed for students who participated in the speaking assessment. This questionnaire consists of six questions: 1) sex; 2) racial/ethnic background; 3) language first spoken at home; 4) whether the student transferred into this school since Christmas vacation; 5) whether the student has been given work in school aimed at improving speaking skills; and 6) how often the student has a chance to speak in classes. A copy of this questionnaire is provided in Appendix F.

## MATERIALS DEVELOPMENT RECOMMENDATIONS

Based on the trial implementation, the following changes are recommended regarding the development and refinement of training materials.

1. Because the examples for the Teacher Observation Training audiotapes were selected from students' actual responses to classroom tasks, the quality of the recordings was necessarily limited. The recording quality of the examples could be improved by employing professional readers who would work from written scripts in a recording studio. This would ensure high quality examples for the training tape.

2. At the minimal level of revision for the Teacher Observation Training audiotapes, six new examples could be selected from the existing tape pool to replace the six examples judged to be of poorest audio quality.

3. Because the content dimension is specifically tied to the nature of the tasks and requires a more lengthy explanation than the other three dimensions, the order of presentation of the dimensions could be changed to delivery, language, organization and content. Reorganization of the dimensions would affect both the written and the taped materials. (See One-on-One Training Session Recommendation No. 3 for related discussion.)

4. Because the first example on the audiotapes was low volume for the delivery dimension, it was difficult for raters to hear and sometimes interfered with their attention to the example. One way to resolve this is to move another example to the beginning of the tapes.

5. For future one-on-one assessments by local school districts, it will be necessary to revise the One-on-One Ratings Supplement so that it addresses the alternate forms of the state speaking tasks.

6. Future one-on-one training sessions would likely use example tapes for the alternate speaking tasks. These could be organized in one of two ways, either by dimension or by task. A complete discussion of the two types of organization is presented in the One-on-One Training Session section and One-on-One Training Session Recommendation No. 3.

## III. TEACHER OBSERVATION TRAINING

The teacher observation phase of the project was designed so that the contractor would train a small cadre of trainers in the teacher observation approach and that these trainers would train teachers in the schools in the use of the observation ratings. The contractor identified three consultants to serve as trainers. These individuals received one day of training as well as a set of materials that they used in their own training from two to three schools.

The training session for trainers was held on April 13, 1981. In the session a complete teacher observation training session was simulated, with the contractor acting as the trainer and the consultants acting as participants. The consultants were given an opportunity to discuss the training materials and strategies. The consultants received all the materials for the subsequent training sessions and worked out logistical problems.

The trainers proceeded to conduct training sessions in all participating schools. Training took place during the period from April 13 to May 1. The size of the training sessions varied and, in some schools, additional teachers who were not participating in the study attended training sessions. In some cases training sessions were as small as seven individuals and in other cases over thirty teachers were trained. Due to scheduling problems, teachers from one school were not all trained in a single session but were split into two sessions.

In the training session, the delivery dimension was presented first, followed by the organization, content and language dimensions. Audiotapes of students' responses to typical classroom speaking tasks were played for each level of each dimension. When there were several characteristics of a particular level, several examples were played. For example, tapes were played to demonstrate poor volume, poor rate and poor articulation, all characteristics of poor delivery. Following the presentation for each dimension, teachers practiced ratings for that dimension. At the end of the training session, teachers practiced rating students responses across all four of the dimensions. Teachers were then asked to rate independently five student responses without discussion. These responses were later compared with the responses of an expert panel and with each other to determine the reliability of the teachers' ratings.

## TEACHER OBSERVATION TRAINING RECOMMENDATIONS

The following recommendations may be made regarding the training of the teacher observation trainers and the training of the teachers in the schools:

1. Possibly the order of training should be changed so that delivery is not presented first. However, there are counter arguments for placing delivery first because it is the easiest dimension in which to be trained.

2. Care should be taken that all participating teachers understand the purposes of the training and the assessment prior to training and express a willingness to participate.

3. Every attempt should be made to conduct future teacher observation training sessions in rooms with adequate acoustics. This will ensure that the tapes are played in the optimum of conditions.

## IV. ONE-ON-ONE TRAINING

The One-on-One Training Session was held on May 8, 1981. All sixteen one-on-one raters attended the four and one-half hour session.

The morning session was devoted to the One-on-One Ratings Supplement. The delivery dimension was presented first, followed by the organization, content and language dimensions. Audiotapes of students' responses to the four speaking tasks were played for each level within the dimensions. This allowed for fewer taped examples to be produced but may have interfered with the raters' distinctions among characteristics of the tasks.

The afternoon session consisted of raters practicing rating students by listening to audiotaped responses to the four speaking tasks. Raters independently made their ratings, the ratings were then placed on wall charts and raters discussed the reasons for their ratings.

Following the practices, raters were asked to rate independently 16 audiotapes without discussion. The tapes consisted of four students responding to all four tasks. From these ratings, the degree of inter-rater reliability was computed.

Finally, raters observed a one-on-one administration in which a student volunteered to be tested before the group. The administration was conducted following the procedures from the Test Administrator's Manual and the student spontaneously responded to the four tasks. Raters were then given the opportunity to clear up any remaining questions regarding the test administration procedures.

### ONE-ON-ONE TRAINING RECOMMENDATIONS

Based on the trial workshop, the following changes are recommended for the One-on-One Training Session:

1. Every attempt should be made to conduct future one-on-one training sessions in rooms with adequate acoustics. This will ensure that the tapes are played in the optimum of conditions.

2. Small group training sessions should be encouraged so that all raters may hear the tapes and be given opportunities to raise questions or concerns.

3. It is suggested that the presentation of the One-on-One Ratings Supplement be structured around the four types of tasks, with the four dimensions discussed within each task. In addition, raters should practice ratings and become reliable for each dimension before moving to the next type of task. For example, the description task could be presented first. Raters would listen to tapes which exemplify

the poor, satisfactory and good levels of the first dimension for the description task. Raters would practice ratings for that dimension for the description task and become reliable before moving to the next dimension. The remaining three dimensions would be presented as they relate to the description task. Raters would practice and become reliable on each dimension, one at a time. This sequence could be repeated for the remaining three types of tasks. This structure would allow the raters an earlier opportunity to practice ratings, which was one of the most valuable parts of the One-on-One Training Session It would require, however, that additional example and practice tapes be selected or developed.

4. Because the content dimension is tied specifically to the nature of the tasks, it is suggested that it be presented as the last dimension. A possible reordering of dimension presentation could be delivery, language, organization and content. This would allow raters to become familiar with the tasks and other dimensions prior to the content dimension.

5. In order to allow raters more time to practice ratings and become reliable for each dimension, one at a time, a longer training session would be advisable. A full-day, eight-hour session could be held or this could be split into a two-day session.

6. It is possible that the characteristics for the poor and good levels of the dimensions could be presented first. The satisfactory level would then be presented as an absence of these characteristics.

7. Inter-rater reliability for the training session was computed by calculating the percentage agreement among the ratings. Reliability of future raters could be calculated against a fixed standard, that is, compared to the "correct" ratings for the tapes prejudged by an expert panel.

## V. DATA COLLECTION FOR THE TEACHER OBSERVATIONS AND ONE-ON-ONE ASSESSMENTS

Upon completion of the Teacher Observation Training Session, four teachers from each school independently rated each student chosen to participate in the assessment. Teachers were instructed to spend three days informally observing the speaking performance of the students they were assigned to rate. Based on the observations and the teachers' recollections of students' performance over the semester, teachers completed one ratings form for each student. The completed ratings forms were then returned to the person assigned as the assessment coordinator for the school. All of the observation ratings were completed during a four-week period from mid-April to mid-May, 1981.

One-on-one assessments were conducted during a two-week period in mid-May, 1981. The one-on-one assessments required approximately two days for each school, with 30 students being assessed each day. Testing took approximately five minutes for each student. Two one-on-one raters independently rated the student's responses to the four speaking tasks using the speaking ratings scales. One-on-one raters took turns giving the tasks to the students. Raters were instructed to follow the procedures presented in the One-on-One Test Administrator s Manual. Raters were asked not to discuss their ratings with one another until completion of the two-day testing period.

### DATA COLLECTION RECOMMENDATIONS

1. Some teachers suggested that they needed more time than three days to do an adequate job of rating students. The teacher observation rating period could be expanded to a longer time frame.

2. Care should be taken so that teachers are assigned to rate only those students who are currently enrolled in their classes. If misassignments occur, students should be reassigned to a current teacher for rating.

## VI. ANALYSIS OF WORKSHOP DATA

At the end of each training workshop, teachers completed a questionnaire that provided information concerning the sex and primary teaching area of the respondents, as well as whether they had taken a course or workshop in teaching oral communication skills. The teachers then listened to five samples of speech and judged them as "poor," "satisfactory," or "good" in each of the dimensions of delivery, organization, content, and language. These data were used to provide an overview of the accuracy of the teachers who would be then going to the schools to judge the students in the study.

Each of the five samples of student speech also was reviewed by a panel of experts. They also judged each student as "poor," "satisfactory," or "good" on each of the four dimensions. A first analysis was undertaken to determine which teachers did not agree with the panel of experts, and to determine whether those discrepancies could be accounted for by the other variables collected.

The first immediate finding was that while the panel of experts tended to rate the students as "satisfactory," the ratings made by the teachers tended to be more extreme. For example, Student 1 was rated as satisfactory in all four dimensions by the panel of experts. However, the ratings assigned by the 97 teachers were as follows:

Table 1
Ratings of Student 1 Made by 97 Workshop Teachers,
Reported by Dimension

| | Number of Teachers Rating Student 1 as: | | |
|---|---|---|---|
| Dimension | Poor | Satisfactory | Good |
| Delivery | 35 | 61 | 1 |
| Organization | 50 | 46 | 1 |
| Content | 45 | 46 | 6 |
| Language | 58 | 38 | 1 |

Notice that while there is a fair degree of consistency among the teachers (the "good" category was almost universally rejected by them in judging this student), their ratings are far from the average of satisfactory determined by the panel. In fact, in two of the four dimensions, a majority of teachers judged the student as "poor." There are two plausible explanations for this disagreement, and both reasons probably had some impact on it. The first is that there was not complete agreement between the experts and the teachers at the end of the workshop as to where "satisfactory" ended and "poor" began. The other is that there was not complete agreement among the experts either, and when there was disagreement, the panel tended to select a rating of "satisfactory".

As a consequence, there was little agreement between the ratings made by the experts and those made by the teachers. Table 2 shows the number of times teachers agreed with the experts on the four ratings made on each student.

Table 2
Number of Times Teacher Ratings Agreed with Expert Ratings, Reported by Student

| | Number of Agreements | | | | |
|---|---|---|---|---|---|
| Student | 0 | 1 | 2 | 3 | 4 |
| 1 | 19 | 15 | 26 | 24 | 13 |
| 2 | 1 | 27 | 35 | 29 | 5 |
| 3 | 2 | 17 | 36 | 32 | 10 |
| 4 | 1 | 30 | 29 | 33 | 4 |
| 5 | 8 | 26 | 37 | 20 | 6 |

It was unusual to find a teacher who agreed with all four ratings made by the panel of experts on any student. Typically, the teachers agreed with the experts on just two of the four content areas, and this was true for all five students. In order to determine whether the teachers and experts agreed more on some dimensions than others, the data in Table 3 were tabulated.

Table 3
Number of Times Teacher Ratings Agreed with Expert Ratings, Reported by Dimension

| | Number of Agreements | | | | | |
|---|---|---|---|---|---|---|
| Dimension | 0 | 1 | 2 | 3 | 4 | 5 |
| Delivery | 0 | 6 | 16 | 35 | 35 | 8 |
| Organization | 9 | 28 | 34 | 16 | 8 | 5 |
| Content | 0 | 11 | 31 | 31 | 21 | 3 |
| Language | 4 | 10 | 31 | 33 | 18 | 1 |

Again, few teachers agreed with the ratings of the experts on all five students, but agreement was noticeably higher in delivery and lower in organization. That result might be due to idiosyncracies in the samples chosen (that is, most of the subjects really were just satisfactory in their delivery, and the one that had good delivery was very good), because deliver is an easier area in which to reach agreement, or because information on how delivery was to be judged was communicated more effectively in the workshop.

On the total of twenty judgments made by each teacher (four judgments each on five students), there were typically between 10 or 11 agreements with the experts. Virtually all (92 percent) had between 7 and 14 agreements, with extremes of 4 and 18 agreements.

Thus, while the original intent of this portion of the study had been to identify the few teachers who did not match the experts' ratings and to see if it were possible to determine why they failed to match, the results nullified this effort. There were so many discrepancies between the experts and the teachers that further pursuit of these data as originally intended was bound to be fruitless.

However, while the teachers did not agree with the experts, even a cursory review of the data revealed that they were in substantial agreement among themselves. Consequently, the analysis was revised to see if it was possible to determine reasons why some teachers assigned ratings that were disparate from the other teachers, rather than from the experts.

A more careful look at the data revealed just how much agreement there was among the teachers. Their total ratings on each of the students are shown in Table 4. Total scores for a student were obtained by summing the teachers ratings (on a scale from 1 to 3) across the four dimensions.

Table 4
Total Scores Assigned by the 97 Workshop Teachers
to Each of the Five Sample Students

| | Sample Student | | | | |
|---|---|---|---|---|---|
| Total Score | 1 | 2 | 3 | 4 | 5 |
| 4 | 18 | 0 | 0 | 29 | 17 |
| 5 | 15 | 0 | 0 | 28 | 36 |
| 6 | 26 | 0 | 4 | 31 | 20 |
| 7 | 18 | 1 | 6 | 5 | 12 |
| 8 | 13 | 5 | 19 | 3 | 9 |
| 9 | 6 | 10 | 25 | 0 | 2 |
| 10 | 1 | 24 | 18 | 0 | 1 |
| 11 | 0 | 30 | 15 | 1 | 0 |
| 12 | 0 | 27 | 10 | 0 | 0 |
| Mean | 6.2 | 10.6 | 9.4 | 5.3 | 5.7 |
| Standard Deviation | 1.5 | 1.2 | 1.6 | 1.2 | 1.4 |

Given the differences in means assigned to the highest and lowest scoring students, the standard deviations are quite small. Also, given the fact that, in the judgment of the experts, the speech of these students did not differ greatly among them (the total expert ratings for the five students were, respectively, 8, 9, 9, 7, and 7), there is great reliability in these judgments.

Another way of showing this reliability is to display the number of teachers who ranked one student higher than another. These results are shown in Table 5.

Table 5
The Number of Times One Sample Student Was Scored Higher Than Another by the 97 Workshop Teachers

| Sample Student | Sample Student 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | - | 95 | 89 | 19 | 31 |
| 2 | 1 | - | 20 | 0 | 0 |
| 3 | 5 | 63 | - | 0 | 1 |
| 4 | 57 | 96 | 94 | - | 43 |
| 5 | 50 | 96 | 92 | 30 | - |

Note: The table entries are the number of times the student listed at the column head was rated higher than the student listed at the row head. For example, Student 2 was rated higher than Student 1 by 95 of the 97 teachers.

The results show some remarkable consistencies. For example, 96 out of 97 teachers assigned higher ratings to Student 2 than to Student 4 or Student 5. Again, given that in the judgment of the experts these students were not that different, this degree of consistency is high. Note also that the teachers ranked the students in the same order as the panel -- Student 2 and 3 as highest, and 4 and 5 as lowest. They simply judged them as being further away from average than the experts did.

The analyses conducted above were rerun by dividing the 97 teachers into two groups -- the fourteen who would conduct the one-on-one observations, and the remaining 83 who would serve as teacher observers. While both the mean and reliability of the fourteen one-on-one raters were slightly lower than the comparable statistics for the 83 teacher observers, neither difference approached statistical significance.

The attempt to use the variables describing the teachers to predict those who would be poor raters proved to be fruitless. As noted earlier, the teacher ratings were so different from the expert ratings that the discrepancies could not be used to judge a teacher as an outlier. Instead, outlying teachers were defined as those whose ratings were different from those of other teachers. To define these points, a rather crude statistic was devised. Outliers were defined to be those whose score on any one student placed them in the extreme 20 scores, in either direction, assigned to a student. Thus, for Student 1 the outlying raters were those who assigned a total score of 4, or a total score of 8 or above; for Student 2, a total score of 9 or below; for Student 3, 7 or below or 12; for Student 4, 7 or above; for Student 5, 8 or above or 4.

The number and direction of outlying scores were tabulated for each teacher. Groups were formed and summary statistics on each of the independent variables were computed. The following results were noted:

1. Sex of the teacher was unrelated to both the direction and size of ratings.

2. The primary course taught by the rater also was unrelated, with one exception: Science teachers tended to have fewer outliers, but also to have fewer positive outliers (i.e., they tended to assign lower scores). However, since there were only 13 science teachers in the study this result may have been an idiosyncracy of the small sample size.

3. Whether the teacher had taken a course or workshop in teaching oral communication was unrelated to the ratings assigned.

4. Teachers in three schools tended to be negative; those in two others were positive. Teachers in three schools had noticeably fewer outliers than average.

In summary, with the exception of whether a teacher taught science, the only variable that appeared related to a tendency to produce outliers was school. As will be noted later in this report, there clearly seemed to be some schools in which the process worked far more satisfactorily than others.

## VII. ANALYSIS OF THE FIELD DATA

After completion of the workshop, the teachers began rating students in their own schools. Some conducted the teacher observations -- the ratings in four dimensions of a student by a teacher. The remainder conducted one-on-one assessments -- the ratings by teachers on four dimensions in four structured speaking situations. The purpose of this section is to report on the results of analyzing data provided from these field settings.

Reliability of the one-on-one raters. In each school, two one-on-one raters observed each student speak in response to four different stimuli (tasks), and rated the student's delivery, organization, content, and language on each task. The purpose of this section is to report on the reliability of those data.

In each school, the correlation between scores assigned by each rater was computed for each data element (a data element is a score on a dimension on a task, e.g., delivery on task 1, or content on task 3). The same correlations also were computed by collapsing across schools.

When each of the four tasks was scored separately, the dimension with the highest reliability on every task was content, and the lowest was language. When a total score for each task was obtained, no task was especially more reliable than any other.

While the reliabilities for any individual task or dimension were not especially impressive (collapsed over schools, they ranged from .33 to .67), the reliabilities of total scores (summing all sixteen scores for each student) were quite high, ranging from a low of .60 in one school to a high of .93 in another. Collapsed over all schools, the reliability of the raters on total score was .78.

While a correlation of this size is not particularly high when compared to reliabilities from paper and pencil tests, it is reaching a fair level of respectability. However, when combined with data from the workshop it becomes quite impressive.

The workshop data that finish the story are the ratings that the one-on-one raters made at the end of the workshop. It turned out that those data were excellent predictors of which pairs would be reliable later on. Data were available from seven pairs -- the one-on-one raters in one school did not assess the five samples at the end of the workshop. The end-of-workshop data were analyzed several ways. The best predictor statistic was also the simplest -- the sum of absolute differences. The results are summarized in Table 6.

Table 6
Descriptive Statistics of the Consistency
of Judgments Made by Pairs on One-on-One Raters

| | End-of-Workshop Data | | | Field Data |
|---|---|---|---|---|
| School | $\Sigma\|D\|$ | $\Sigma D^2$ | r | r |
| A | 4 | 6 | .74 | .90 |
| B | 5 | 11 | .90 | .93 |
| C | 6 | 8 | .93 | .89 |
| D | 6 | 10 | .72 | .81 |
| E | 7 | 16 | .79 | .72 |
| F | 9 | 19 | .42 | .69 |
| G | 9 | 35 | .59 | .60 |

Note: $\Sigma|D|$ is the absolute differences between the two raters summed over the five workshop sample students.

$\Sigma D^2$ is the square of the differences between the two raters summed over the five workshop sample students.

r is a Pearson correlation coefficient between the scores assigned by the two raters.

Thus, the data suggest that simply by having the two raters rate each of the five samples, obtaining a total score for each sample for each rater, determining the differences, and then adding them, one obtains a measure that forecasts fairly accurately how reliable those judges will be. Given how similar the two tasks are, it does not seem that surprising in retrospect that one should predict the other so well.

Now, however, suppose that rather than having all pairs conduct one-on-one ratings, data had been collected only from those pairs whose sum of absolute differences on the five sample students were less than or equal to 6. The reliability of those four pairs, collapsed across schools, turned out to be .86. This is a level of reliability that should be more than acceptable for an assessment of this nature.

It was also worthwhile noting that while no task appeared to be especially more reliable than any other when the data were collapsed over all eight schools, this was not the case in the four high reliability schools. In all four of these schools, task 4 was the most reliable, and in three schools, task 3 was the least reliable.

Before leaving this topic we will provide one additional illustration of how strong the relationship is between the scoring exercise at the end of the workshop and the actual field practice of conducting a one-on-one assessment. From the scoring exercise, the reliability of the one-on-one raters was determined to be .85 for five ratings. This implied the reliability would be .82 for four ratings. This predicted reliability turned out to be almost the exact reliability observed in the field.

Reliability of the teacher observers. In each school, students were rated by four teachers, who evaluated the students' speaking level on a scale from 1-5 in four areas: delivery, organization, content, and language. Teachers also assigned an overall rating. The purpose of this section is to report on the reliability of those data.

In short, the data correlating one teacher with one other teacher were not especially reliable. The reliability of the dimensions ranged from .35 for delivery to .47 for language. The average correlations for both the sum of scores and for the "overall" category assigned by the teachers was .47.

In two schools, however, the reliability of the teacher observation data was distinctly higher -- .77 in one school and .70 in the other. Unfortunately, however, none of the data collected at the workshop would lead one to predict such a result.

When the results for two teachers were summed and correlated with the sum from another pair, the reliabilities naturally increased. This statistic ranged from .50 in the least reliable school to .86 in the highest. However, the doubling of the number of observers did not increase the reliabilities to the extent that would be predicted by the Spearman-Brown prophecy formula in six of the eight schools. This implies that the two additional teachers were not judging the student on quite the same basis as the first two. This is not unexpected, but also means that it is unlikely that the unreliability problem of the teacher observation data can be solved merely by requiring more teachers to judge each student.

One result which was somewhat surprising was that the second pair of teachers assigned lower scores to the students than did the first pair. The difference was over one point. Given that the standard deviation of these total scores was less than 6, this difference is a noticeable amount. The result was due to the fact that teachers who indicated their primary teaching area as "social studies" or "other" scored students about a point below teachers from the other three areas. Since social studies and "other" teachers usually were in the second pair, the ratings from that pair were lower than those from the first pair.

## VIII. COMPARING THE TEACHER OBSERVATION MEASURE TO THE ONE-ON-ONE MEASURE

The reason for investigating two measures was the anticipation that one could be used as a proxy for the other. More specifically, it was hoped that the teacher observations could be obtained for all students. Then, only those students who fell in some borderline category would be assessed by a one-on-one rater.

An underlying assumption of this process is that, except for reliability issues, the two measures are equivalent. Thus, the next question to be addressed dealt with the equivalence of the two measures.

The data indicated that the two measures were not equivalent on a variety of bases. First, the mean score assigned by the one-on-one raters (on their scale of 16 to 80) was 48.8, while the mean score of the teacher observers (which should have been half that of the one-on-one raters, given the scale of 8 to 40 that the teacher observers were using) was 26.0. Thus, the difference in mean ratings on these two scales, when adjusted to make them comparable, was about a third of a standard deviation.

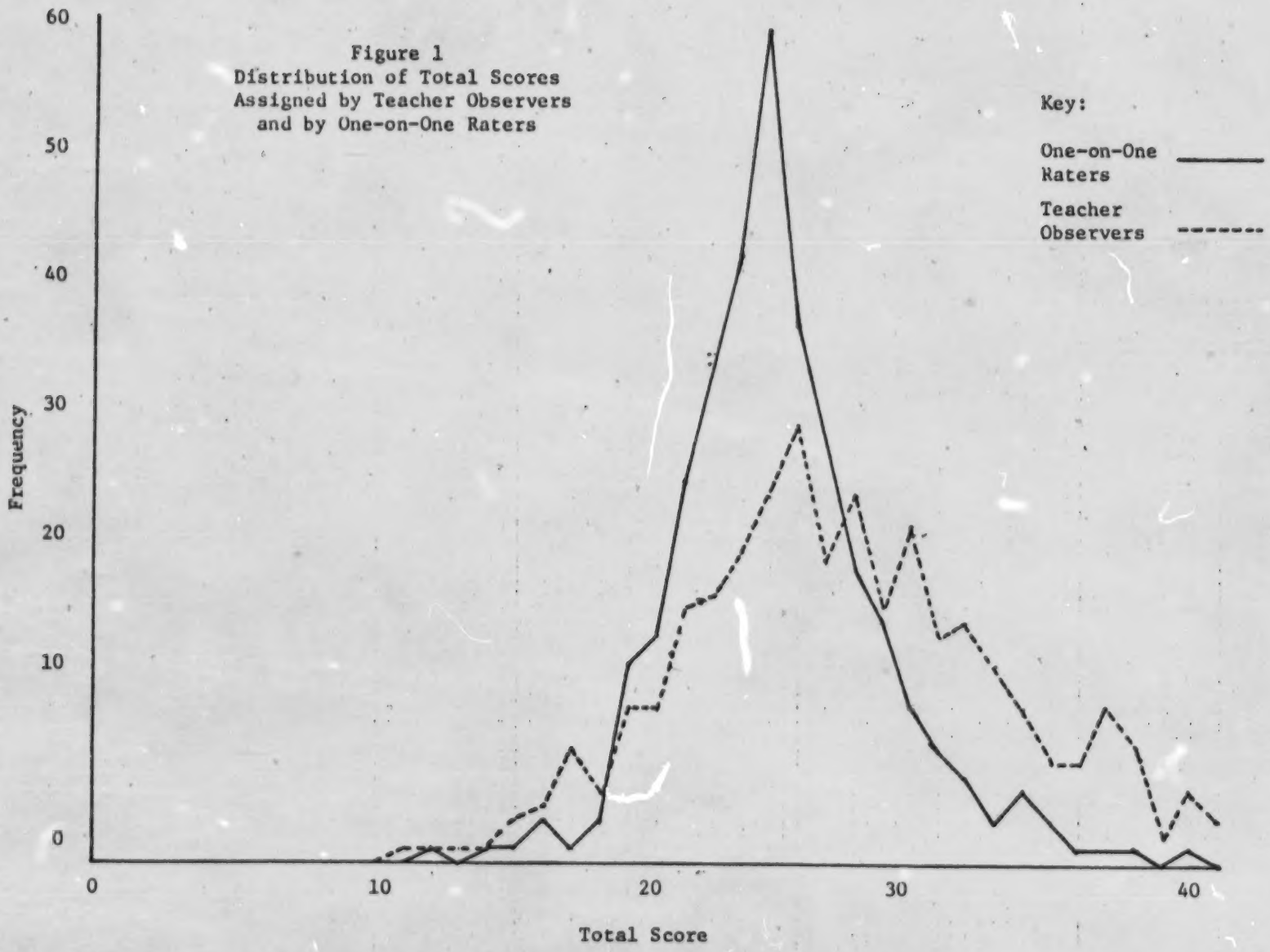
Secondly, the variances of the two scales were very different. The variance of the one-on-one ratings was about half what it needed to be to be equivalent to the teacher observations. Figure 1, located on page 20, shows the distribution of scores for 351 students who were scored by four teachers and two raters (the one-on-one ratings are divided by two to place them on the same 8-40 scale as the teacher observations). The difference in mean and variance between the two sets of scores is obvious from the graph.

Generally, measures that are more variable are more reliable. However, this was not the case here. The more variable teacher observation measure had poor reliability. The correlation across all schools between the sum of ratings assigned by one pair of teachers and the sum assigned by the other pair was .65. In one school, the reliability of the teacher observations measure was .86, but in half the schools it was lower than .60.

In contrast, the reliability of the one-on-one ratings was .78 overall. In two schools, the reliability was .90 or higher, and above .80 in two more schools. More importantly, it was possible to predict from the workshop results which schools were going to have the highest reliabilities.

In an earlier section, the ratings that raters completed at the end of each training session were described. The differences between the ratings assigned by the two raters in a school were computed for each sample, and the absolute values of these differences were summed over the five samples for each pair of raters. In four schools, this sum of differences totaled 6 or less. Those were the same four schools that had one-on-one reliabilities over .80. The overall reliability of these four schools

Figure 1
Distribution of Total Scores
Assigned by Teacher Observers
and by One-on-One Raters



was .86. Thus, it was possible, on the basis of a preliminary measure, to identify a set of raters who would produce satisfactory reliability.

In contrast to the high predictability and reliability of the one-on-one measure, the teacher observation data generally had low reliability. In addition, in those few cases where the teacher observation data were reliable, there was nothing in the background data to predict such reliability. Several analyses were done to attempt to isolate any factors that might have led to higher predictability. The reliability of the teachers on the workshop samples was computed for each school. The school that had the highest reliability on the workshop samples had the lowest reliability for the field teacher observation data; and the school that had the lowest reliability on the workshop samples had the highest reliability for the field teacher observation data.

The reliabilities of the teachers in the various subject matter areas were computed. They are shown in Table 7.

Table 7
Correlations Between Observation Ratings
by Teachers in Various Subject Matter Areas
and One-on-One Ratings

| | Math | Science | Social Studies | Other | One-on-One Ratings |
|---|---|---|---|---|---|
| English | .63 | .63 | .64 | .34 | .35 |
| Math | -- | .59 | .60 | .34 | .19 |
| Science | | -- | .54 | .44 | .37 |
| Social Studies | | | -- | .43 | .40 |
| Other | | | | -- | .2[illegible] |

The results show that the subject matter area of the teacher had little effect on the ratings, with the exception that teachers not in one of the four main categories consistently had the lowest correlations. Science and social studies teachers had higher correlations with the one-on-one ratings than did the English and math teachers, a result that was somewhat surprising. And most important to later discussion, the teachers consistently correlated more highly among themselves than they did with the one-on-one raters.

In a final attempt to extract some pattern of predictability from the results, a reliability coefficient was computed for every teacher by computing across all observations made by that teacher, the correlation between that teacher's ratings and:

a) the sum of the ratings made by the other three teachers

b) the ratings made by the one-on-one raters

c) a weighted average of the other three teachers and the two one-on-one raters.

These correlations ranged from a low of -.81 to a high of 1.00. Of course, these extremes were registered by teachers who had provided observations on only a few students. However, seven teachers with between 57 and 60 observations each had correlations with the sum of the other three observations of .80 or higher. No one with any substantial number of observations had correlations with the one-on-one ratings above .67 with either of the two raters, however.

The workshop data for the teachers with the highest 20 and lowest 20 correlations with the other three teachers were compared to see if there were differences between the two groups. There were none observed, except that those with higher correlations had tended to err on the high side in judging the workshop samples, while those with low correlations tended to err on the low side. Also, there were three schools that had all their students judged by just four teachers. From these three schools came the two highest reliability coefficients. However, the third had a reliability of only .58. And in none of these schools was the correlation with the one-on-one ratings especially high.

In short, the data all seem to indicate that the measure obtained from the teacher observers is quite different from that obtained from the one-on-one raters. The means, variances, reliabilities and predictabilities of the measures are all different. Additionally, the correlation between the two, even after correcting for unreliability, is something less than .70. This figure compares unfavorably to the correlations of .80 and higher typically found in different content areas of standardized achievement test batteries. Consequently, it appears unjustifiable to combine these two measures. A decision must be made as to which is the valid one, for given the low correlation between the two, if one is highly valid the other cannot be.

## IX. CHOOSING BETWEEN THE TWO MEASURES

Needless to say, the statistical analysis of these measures will not be the only determinant of which one ultimately will be judged to be the more valid. However, some observations about them can be justifiably made at this point.

1. The one-on-one ratings are made under standardized conditions; the teacher observations are influenced by the varied conditions under which teachers have observed their students.

2. The one-on-one ratings are more highly reliable than the teacher observations. Also, it appears that there is some predictability as to how reliable those ratings will be.

3. Though the teacher observations usually are not reliable, they sometimes are. And since two observations are made on each student, an estimated reliability coefficient always can be computed. For example, in one school, the correlation between teacher 1 and teacher 2 was .80. Using the Spearman-Brown prophecy formula, we would predict the reliability of the sum of those two scores to be .89. The actual correlation between the first pair of raters in that school and the second pair turned out to be .86. Thus, the high reliability in that school could have been determined from the data collected to assess the students. Consequently, one possible approach might be to allow schools to attempt to rate students, observe the correlations obtained, and then make a decision as to whether to discard the process because it is unreliable or accept it because the reliability is sufficiently high.

If the one-on-one rating procedure is deemed to be the more valid approach to assessing speaking, it may be desirable to increase its efficiency. A natural question is whether its efficiency can be increased without seriously affecting its reliability. To answer this question, two stepwise regression analyses were run. In each analysis, the score obtained by summing the four task scores for one rater was used as the dependent measure, and the four task scores of the other rater were used as the independent measures. The results of both analyses were virtually identical, with task 4 loading first and task 1 entering second. The correlation between the sum of the two tasks by one rater and the sum of all four tasks by the other rater was .75 going one way and .77 going the other. The correlation of the sum of the two tasks and the sum of the four tasks for the same rater was .95 for one rater and .94 for the other. The reliability of the sum of the two tasks over all schools was .79, which compares favorably to the reliability of .78 found for the sum of all four tasks. More importantly, the reliabilities for each school held relatively constant when just tasks 1 and 4 were summed, as

shown in Table 8. Thus, one can conclude that eliminating tasks 2 and 3 from the assessment would have little effect on the reliability of the process. However, this may affect the validity of the measure.

Table 8

The Reliability of the One-on-One Raters, Using a Total Score on All Four Tasks as a Measure, and Using a Total Score on Tasks 1 and 4 as a Measure, Reported by School

| School | Summing All Four Tasks | Summing Task 1 and Task 4 |
|---|---|---|
| 1 | .93 | .93 |
| 2 | .72 | .73 |
| 3 | .60 | .65 |
| 4 | .90 | .87 |
| 5 | .77 | .76 |
| 6 | .81 | .77 |
| 7 | .69 | .74 |
| 8 | .89 | .90 |
| Overall (Collapsing across all schools) | .78 | .79 |

The current rating scale has five points. One potential way of simplifying the rating would be to collapse the first two and last two points on the scale; i.e., to make the scale have three points rather than five. When the one-on-one data were rescored on the reduced scale, the results were surprising. The shrinking of the scale had minimum impact when all four tasks were summed: the correlation of the scores on the reduced scale with those on the full scale was .98, and the reliability of the reduced scale was .77, as opposed to .78 for the full scale. But when only two tasks were scored on the reduced scale, there was substantial impact: the correlation of the scores on the reduced scale with those in the full scale was only .92, and the reliability is affected. When four tasks are summed on the scale from 1 to 5, there are 65 points on the scale (scores can range from 16 to 80). When two tasks are scored on the full scale, or when four tasks are scored in the reduced scale, there are 33 points on the scale. But when just two tasks are scored on the reduced scale, there are only 17 points on the scale, and this apparently is too few for reliable measurement. (Parenthetically, there are only 17 points on the teacher observation scale. Could this be a reason for its lower reliability?).

It must be noted that the above results were obtained by shrinking the scale after the raters had made their judgments on the full scale. What the reliabilities would have been had they started with a reduced scale is a matter for speculation at this point.

## X. RECOMMENDATIONS

1. A decision must be made as to which measure is going to be used to measure speaking. Since the two intercorrelate so poorly, any system which uses both in tandem would be illogical. One is not a more or less reliable substitute for the other; they are two different measures.

2. The one-on-one rating measure appears to be the measure whose validity can be better justified. The standardization of the observations is a major reason for recommending selection of one-on-one ratings as more valid. Only if it can be shown that the tension and/or artificiality associated with the standardized setting seriously erodes its validity should teacher observations be chosen as the preferred measure.

3. If teacher observations are chosen as the preferred measure, checks of reliability would have to be made after the collection of data. Given the results of this study, it is anticipated that these reliabilities would be unsatisfactory in a vast majority of schools, and there would be no way of predicting before the data were collected which schools would prove to be insufficiently reliable. Consequently, much data would be collected that would ultimately turn out to be of unsatisfactory quality.

4. The one-on-one ratings have the advantage of being reliable, and this reliability is somewhat predictable. An effort should be made to increase this predictability, for several reasons:

   a. Because it is anticipated that only one rater will be judging each student, there is no way of determining a rater's reliability after the fact. The reliability of the rater must be known from some other source.

   b. While many raters will be sufficiently reliable, some will not. If these less reliable raters do much judging, the overall reliability of the entire process could be lowered.

   c. Even if reliability could be checked after the fact, one would want to discard as little data as possible. Having raters in the field with high reliability would minimize this problem.

The effort to increase predictability should focus on having a sufficient number of samples to be rated at the end of the workshop. Only those raters who provide scores for those samples within a tolerable range should be approved to conduct one-on-one ratings in the schools.

5. If valid to do so, the number of speaking tasks should be reduced to three, and possibly two. It appears that tasks 2 and 3 could be eliminated from the test without affecting the reliability of the scoring.

6. The Department should focus its energies on finalizing the one-on-one rating process. It should demonstrate that the measure is valid and can be scored reliably, that reliability can be maintained at a high level by requiring raters to pass a reliability test and that the situations presented and the ratings made are done without bias. In order to accomplish these outcomes, the Department should sponsor a study that includes the following elements:

a. The training of a large number of potential raters, divided equally among the major racial/ethnic groups; about sixteen raters of each racial/ethnic group would be a desirable number. Before this training began, the knowledge gained from this year's study should be used to revise the training materials and manuals.

b. The reliability testing of these raters after training on their ratings of an audio taped selection of eight to ten students.

c. The ratings by these raters of a group of twelve students. Again these students would be divided equally among the major racial/ethnic groups. These students would be rated simultaneously be all raters -- perhaps using videotapes, perhaps live.

d. The raters would be paired and conduct joint ratings of thirty students per pair in schools convenient to them.

The data from this study could be used to determine the validity of the reliability test as well as differential scoring biases that might exist when raters and examiners are from different ethnic groups. Such a study would provide the necessary information to determine whether the one-on-one ratings could be used confidently in the schools, with the knowledge that they are valid, reliable and free of bias.

# APPENDIX A

## TEACHER OBSERVATION RATINGS GUIDE
## ONE-ON-ONE RATINGS SUPPLEMENT

MASSACHUSETTS DEPARTMENT OF EDUCATION

ASSESSMENT OF BASIC SKILLS

TEACHER OBSERVATION RATINGS GUIDE

## OVERVIEW

There are numerous kinds of speaking tasks a student must perform in the classroom. These include such things as asking questions, answering questions, presenting reports and participating in group discussions.

In order to accomplish a particular speaking task, the student must formulate and transmit a message to one or more listeners. This process involves deciding what needs to be said, organizing the message, adapting the message to the listeners and the situation, choosing language to convey the message and, finally, delivering the message.

The Massachusetts Assessment of Basic Skills in speaking focuses on four dimensions which represent the major components of the speaking process:

- Delivery
- Organization
- Content
- Language

Delivery is concerned with the transmission of the message, i.e., volume, rate and articulation. Organization is concerned with how the ideas in the message are related to one another and how the ideas are ordered. Content is concerned with the amount and relevance of information in the message and how the content is adapted to the listeners and the situation. Language is concerned with the grammar and words that are used to convey the message.

The student is rated on each of the four dimensions using a 5-point scale: 1 is the lowest rating and 5 is the highest rating. The specific definitions for the various points of the scale are different for each dimension. However, across dimensions, the various points of the scale follow the same pattern. This pattern takes into account both the level of the student's performance and the consistency of the student's performance.

In a particular speaking task, the student demonstrates one of several levels of speaking performance. These levels may be labeled poor, satisfactory and good. Poor performance is characterized by speaking behaviors that detract from the communication of the message and hinder understanding of the message. Satisfactory performance is characterized by speaking behaviors that meet the requirements of the speaking task but do not exceed them. Good performance is characterized by speaking behaviors that enhance the communication of the message and help understanding of the message.

At different times throughout the school year the student performs at different levels. In one task the student may exhibit good speaking performance, while in another task the student may exhibit only satisfactory speaking performance. However, over time the student demonstrates some degree of consistency in his or her level of performance. The degree of consistency is reflected in the rating scale. In some cases, the student may be described as performing at one level most of the time. In other cases, the student may be described as performing at one level some of the time and at another level some of the time. In the latter cases, the student usually performs at poor and satisfactory levels or at satisfactory and good levels. It is rare that the student performs at both poor and good levels.

Within each dimension, the student is rated according to the level and the consistency of his or her speaking performance. The points of the rating scale follow the same general pattern:

1 = Poor performance most of the time.

2 = Poor performance some of the time;
satisfactory performance some of the time.

3 = Satisfactory performance most of the time.

4 = Good performance some of the time;
satisfactory performance some of the time.

5 = Good performance most of the time.

The five points of the rating scale are illustrated in the exhibit on the following page.

## LEVELS AND CONSISTENCY OF PERFORMANCE

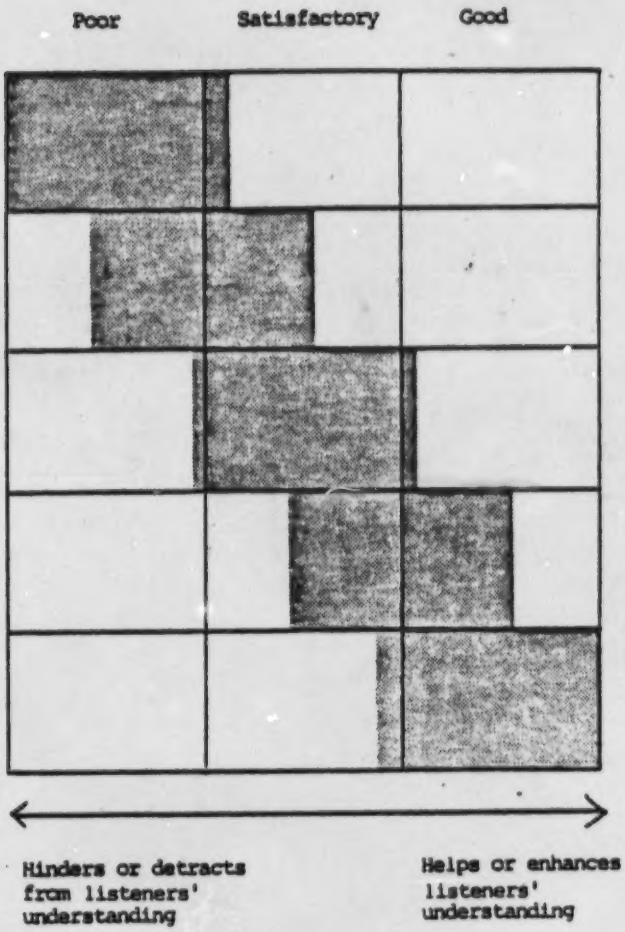
1 Poor most of the time

2 Poor some of the time
Satisfactory some of the time

3 Satisfactory most of the time

4 Good some of the time
Satisfactory some of the time

5 Good most of the time



In evaluating the student, the rater needs to consider three things:

1. The four dimensions of speaking performance
   - Delivery
   - Organization
   - Content
   - Language
2. The general level of performance within each dimension
   - Poor
   - Satisfactory
   - Good
3. The consistency of performance within each dimension
   - Most of the time
   - Some of the time

The rater needs to consider each dimension separately. The student does not necessarily perform at the same level or with the same consistency in each dimension. For example, the student may have poor delivery most of the time but have good content some of the time.

The rater also needs to evaluate the student's speaking performance in an objective way. First, the rater must focus on the speaking performance of the student and not other factors, such as academic performance, deportment or appearance. The rater must be careful not to be influenced by the student's knowledge of the subject matter, general behavior in class, dress or grooming. Second, the rater must not let a particular set of norms of social acceptability influence the rating. The rater must evaluate the student only in terms of how well the student meets the requirements of various speaking tasks. The student's accent or use of non-standard English may or may not be important depending upon the requirements of the task.

Finally, the rater needs to evaluate the student's performance on the average in a variety of situations. It is possible that a student occasionally demonstrates good performance and occasionally demonstrates poor performance. In evaluating the student, the rater needs to consider the student's performance on the average over time and not be overly influenced by occasional instances of extremely good or extremely poor performance. Also, it is possible that the student performs differently depending on the situation. Many students experience a great deal of anxiety in formal speaking situations and do not perform as well as normal in these situations. In evaluating the student, the rater needs to consider the student's performance in a variety of situations and not be overly influenced by performance in certain types of situations, such as speeches in front of the class.

## DELIVERY

The delivery rating focuses on the transmission of the message. It is concerned with volume, rate and articulation. Rate refers to the speed and fluency of speech. Some examples of rate problems include stammering, long pauses and vocalized pauses, such as ahs, uhms and "you knows." Articulation refers to pronunciation and enunciation. Some examples of articulation problems include incorrect pronunciation, mumbling and slurring words.

POOR

The student's delivery is poor if one or more of the following statements is true:

- The student's volume is so low it is hard to understand what he or she is saying. The student's volume is so loud it is distracting.
- The student's rate is so fast it is hard to understand what he or she is saying. The student's rate is so slow it is distracting. The student stammers. The student pauses for a long time or at inappropriate places so that it is distracting. The student fills pauses with ahs, uhms , "you knows" and so forth.
- The student mispronounces words making it hard to understand what he or she is saying. The student mumbles or slurs words.
- You have difficulty understanding what the student is saying because of problems in delivery.
- You are distracted from what the student is saying by problems in delivery.

SATISFACTORY

The student's delivery is satisfactory if most of the following statements are true:

- The student's volume is not too low or too loud.
- The student's rate is not too fast or too slow. The student seldom stammers. The student seldom pauses too long or at inappropriate places. The student seldom fills pauses with ahs, uhms, "you knows" and so forth.
- The student seldom mispronounces words. The student seldom mumbles or slurs words.
- The student's delivery does not hinder your understanding of what he or she is saying.

- You are not distracted from what the student is saying by problems in delivery.

GOOD

The student's delivery is good if one or more of the following statements is true:

- The student varies his or her volume to add emphasis or interest.
- The student varies his or her rate and uses pauses to add emphasis or interest. The student speaks in a lively, enthusiastic fashion.
- The student's enunciation is clear and distinct.
- The student helps you understand the message by using good delivery.
- You notice that the student's delivery is particularly effective.

NOTE: You may wonder if you should give the student a lower rating because the student speaks with an accent. Accent should be rated with respect to a person's ability to understand the message, not the social acceptability of the accent. One particular accent is not better or worse than another. However, if you think that most listeners would have difficulty understanding what the student is saying because of his or her accent, then you should give that student a lower rating in delivery. Also, make sure you are rating the student on his or her delivery on the average in a variety of situations. Keep in mind that many students have more problems with delivery in formal speaking situations than normal and do not be overly influenced by the student's performance in these situations.

REMEMBER, in the delivery dimension you are rating how the student speaks, not what the student says. The ratings for this dimension are:

1 = Poor delivery most of the time.

2 = Poor delivery some of the time;
satisfactory delivery some of the time.

3 = Satisfactory delivery most of the time.

4 = Good delivery some of the time;
satisfactory delivery some of the time.

5 = Good delivery most of the time.

## ORGANIZATION

The organization rating focuses on how the content of the message is structured. It is concerned with the relationships among ideas and the order of ideas in the message. The order of ideas in messages depends partly on the type of speaking task. In some speaking tasks the content usually is organized in a step-by-step order. In other speaking tasks the content is organized with a generalization first followed by details. There are many types of appropriate orders.

### POOR

The student's organization is poor if one or more of the following statements is true:

- The relationships among ideas are unclear. The student moves from idea to idea in a random or rambling fashion. The student fails to provide the transitions or connections between ideas necessary for understanding the message. The student's references to previous statements are vague or misleading.
- The ideas are presented in a mixed up order. For example, the student talks about one idea, then talks about another idea and then jumps back to the first idea.
- You have difficulty understanding what the student is saying because the relationships among ideas are unclear.
- You have to make assumptions about the relationships among ideas or the order of ideas in the student's messages.

### SATISFACTORY

The student's organization is satisfactory if most of the following statements are true:

- The ideas in messages are related to one another.
- The ideas in messages are not presented in a mixed up order.
- Problems of organization do not hinder your understanding of what the student is saying.
- You do not have to make assumptions about the relationships among ideas or reorder the ideas in the student's messages.

GOOD

The student's organization is good if one or more of the following statements is true:

- The student begins statements by naming the topic or by saying what he or she is going to talk about. For example, the student says "This is a story about ...."
- The student uses clear transitions. For example, the student says "The first step is ...", "next ..." and so forth.
- The student completes statements by summarizing what he or she talked about.
- The student helps you understand what he or she is saying by presenting messages in a well organized way.
- You notice that the student is particularly well organized when he or she talks.

NOTE: Make sure you are not unconsciously "filling in" transitions or connections or reordering ideas for the student because you can figure out what the student is talking about. If you have to make assumptions about the relationships among ideas in messages or reorder the ideas, then you should give that student a lower rating in organization.

REMEMBER, in the organization dimension you are rating how the student organizes the message, not what the student says. The ratings for this dimension are:

1 = Poor organization most of the time.

2 = Poor organization some of the time;
satisfactory organization some of the time.

3 = Satisfactory organization most of the time.

4 = Good organization some of the time;
satisfactory organization some of the time.

5 = Good organization most of the time.

## CONTENT

The content rating focuses on what is said. It is concerned with the amount of the content, the relevance of the content and the adaptation of the content to the listeners and the situation. Adaptation involves adjusting messages to the knowledge and experience of the listeners and to aspects of the situation, such as the degree of formality or the number of people present. For the purpose of rating speaking skills, the accuracy of the content is not relevant.

### POOR

The student's content is poor if one or more of the following statements is true:

- The student says very little. The student does not provide enough content to meet the requirements of speaking tasks.
- The student focuses on content that is irrelevant to the speaking task. The student wanders off the topic.
- The student seems to ignore the listeners and the situation when he or she talks. The student uses words and concepts that are inappropriate for the knowledge and experiences of the listeners or for the situation. For example, the student uses terms that the listeners would not be expected to know or uses slang in a formal situation. When the student presents arguments, he or she does not take the listeners into account.
- You have only a vague idea of what the student is talking about.
- You have to add information in order to understand the student's messages.

### SATISFACTORY

The student's content is satisfactory if most of the following statements are true:

- The student provides enough content to meet the requirements of speaking tasks.
- The student focuses primarily on content relevant to the speaking task. The student sticks to the topic.
- The student seems to be generally aware of the listeners and the situation when he or she talks. The student uses words and concepts that are appropriate for the knowledge and experience of most listeners and for most situations. The student uses arguments that are appropriate for most listeners.

- You have a general idea of what the student is talking about.

- You do not have to add information in order to understand the student's messages.

GOOD

The student's content is <u>good</u> if <u>one or more</u> of the following statements is true:

- The student elaborates on the topic when he or she talks.

- The student provides a variety of types of content appropriate for speaking tasks, such as generalizations, details, examples and various forms of evidence.

- The student adapts the content of messages to the specific listeners and situation. The student takes into account the specific knowledge and experience of the listeners, adds explanations as necessary and refers to the listeners' experience. The student uses arguments that are adapted to the specific values and motivation of the listeners.

- You learn a lot about what the student is talking about.

- The student helps you understand messages by adapting the content specifically to you.

NOTE: Make sure you are rating the student on the quantity, relevance and adaptation of the content, not the accuracy of the content. Concerns with the accuracy of the content fall outside a speaking skills assessment. Also, make sure you are not unconsciously "filling in" content for the student because you happen to know something about what the student is talking about. If you have to add information, then you should give that student a lower rating in content.

REMEMBER, in the content dimension you are rating <u>the quantity, relevance and adaptation of what the student says</u>, not the accuracy of what the student says. The ratings for this dimension are:

1 = Poor content most of the time.

2 = Poor content some of the time;
satisfactory content some of the time.

3 = Satisfactory content most of the time.

4 = Good content some of the time;
satisfactory content some of the time.

5 = Good content most of the time.

## LANGUAGE

The language rating deals with the language that is used to convey the message. It is concerned with grammar and choice of words. Grammar refers to the structural rules for sentences, such as noun-verb agreement, verb tense and word order. Choice of words refers to the accuracy and adequacy of the words used to convey concepts and meaning.

### POOR

The student's language is poor if one or more of the following statements is true:

- The student makes so many grammatical mistakes it is hard to understand what he or she is saying.
- The student's oral vocabulary is limited. The student uses simplistic, bland language. The student uses the wrong word to express what he or she means to say.
- You have difficulty understanding what the student is saying because of mistakes in grammar or inaccurate choice of words.
- Your understanding of what the student is saying is limited because he or she uses very simplistic, bland language.

### SATISFACTORY

The student's language is satisfactory if most of the following statements are true:

- The student seldom makes grammatical mistakes.
- The student's oral vocabulary is adequate to meet the requirements of speaking tasks.
- The student's grammar or choice of words does not hinder your understanding of what the student is saying.
- You are not distracted from what the student is saying by problems in language.

GOOD

The student's language is good if one or more of the following statements is true:

- The student's oral vocabulary is extensive.
- The student uses various types of language appropriate for speaking tasks, such as descriptive, persuasive or emotional language, humor, imagery, metaphor and simile.
- The student helps you understand what he or she is saying by using a variety of types of language.
- You notice that the student uses language that is particularly effective.

NOTE: You may wonder if you should give a student a lower rating because the student does not speak English very well or uses a non-standard form of English. Language should be rated with respect to a person's ability to understand the message, not the social acceptability of the grammar or choice or words. However, if you think that most listeners would have difficulty understanding what the student is saying because of problems with speaking English or use of non-standard English, then you should give that student a lower rating in language.

REMEMBER, in the language dimension you are rating how the student conveys the message through language, not what the student says. The ratings in this dimension are:

1 = Poor language most of the time.

2 = Poor language some of the time;
satisfactory language some of the time.

3 = Satisfactory language most of the time.

4 = Good language some of the time;
satisfactory language some of the time.

5 = Good language most of the time.

## DIRECTIONS FOR COMPLETING TEACHER OBSERVATION RATINGS

1. Before you begin rating students, read the Teacher Observation Ratings Guide thoroughly.

2. After reading the guide, spend three days informally observing the speaking performance of the students that you have been assigned to rate.

3. After observing the students, complete one Teacher Observation Ratings form for each student that you have been assigned to rate. Notice that the form is printed on both sides of the sheets you have received so that you may complete ratings for two students on a single sheet. Rate one student at a time. Do not let your ratings of one student influence your ratings of another student.

4. Complete the dimension ratings. Using the 5-point scale, rate each student with respect to his or her delivery, organization, content and language. Base your ratings on your observations of the student's speaking performance over the last three days and on your recollections of his or her performance over the last semester. Be sure to base your ratings on the student's performance on the average in a variety of situations. Do not be influenced by occasional instances of performance that are not typical for the student.

5. Complete the overall rating. Using the 8-point scale, rate each student with respect to his or her overall speaking skills. Base your rating on your impression of the student's speaking skills overall. The overall rating is different from the dimension ratings. You should establish the overall rating independently. Do not derive the overall rating from the numerical average of the dimension ratings.

6. For each student, complete the questions at the bottom of the ratings form.

7. Complete the ratings forms for all of the students that you have been assigned to rate within three school days after your training session. Complete the ratings for students even if you have not had a chance to observe their speaking performance because the students were absent or were not involved in any speaking activities. Return all the ratings as directed below:

Return to: _____

By: _____

MASSACHUSETTS DEPARTMENT OF EDUCATION

ASSESSMENT OF BASIC SKILLS

ONE-ON-ONE RATINGS SUPPLEMENT

## OVERVIEW

For the one-on-one assessment, a specially trained rater rates the student's performance using speaking rating scales similar to those used for the teacher observation approach. However, instead of basing ratings on classroom observations, the one-on-one rater gives the student four specific speaking tasks and rates the student's performance on these tasks.

The speaking tasks used in the one-on-one approach represent four speaking situations that are important for the student's everyday functioning both in and out of school. They include:

- A description task
- An emergency task
- A sequence task
- A persuasion task

For the description task the student is asked to describe a favorite class or extracurricular activity. For the emergency task the student is asked to simulate calling the fire department to report smoke in the house. For the sequence task the student is asked to explain how to cook something. For the persuasion task the student is asked to simulate convincing the principal to make one change in the school.

The scales used to rate the student's performance on each of the four tasks follow the same general pattern as the scales used for the teacher observations. The student's performance is evaluated with respect to four speaking dimensions:

- Delivery
- Organization
- Content
- Language

For each dimension the student is rated using a 5-point scale: 1 is the lowest rating and 5 is the highest rating.

Like the teacher observation rating scales, the one-on-one rating scales take into account the level of the student's performance:

- Poor
- Satisfactory
- Good

At the poor level the student exhibits problems with speaking performance. At the satisfactory level the student does not exhibit major problems with speaking performance but also does not exhibit characteristics of particularly good speaking performance. At the good level the student exhibits characteristics of good speaking performance.

The one-on-one rating scales also take into account the consistency of the student's performance. However, instead of considering the consistency of the student's performance over an entire semester, the one-on-one rating scales take into account the frequency of speaking behaviors that the student exhibits during a single response to a task. In some cases a student may exhibit many problems with speaking performance during his or her response. The problems are pervasive throughout the response. In other cases, the student may exhibit some problems with speaking performance during his or her response. The problems are intermittent during the response.

Within each dimension, the student is rated according to the level and the consistency of his or her speaking performance in response to a task. The points of the rating scale follow the same general pattern:

1 = Many problems with speaking performance

2 = Some problems with speaking performance

3 = Absence of problems with speaking performance and absence of characteristics of good speaking performance

4 = Some characteristics of good speaking performance

5 = Many characteristics of good speaking performance

The rater needs to evaluate the student's performance based on his or her overall impression of the student's response. The rater should not be influenced by infrequent problems with performance or infrequent characteristics of good performance. The rater should only give a student a poor rating if the student exhibits a noticeable number of problems with performance. The rater should only give a student a good rating if the student exhibits a noticeable number of characteristics of good performance.

In some cases a student may exhibit problems in performance and characteristics of good performance within a single response. In these cases, the rater needs to take into account the relative frequencies of these behaviors. For example, if the student has basically good performance with an occasional problem, the student should still be given a good rating. However, if the student exhibits a few problems and a few characteristics of good performance, but the overall impression is of satisfactory performance, then the student should be given a satisfactory rating.

To a great extent, the one-on-one ratings follow the same pattern for all four speaking tasks. However, the characteristics of good performance differ from task to task. For example, good delivery for the description task is likely to be lively and enthusiastic while good delivery for the emergency task is likely to be clear and distinct. All levels (poor, satisfactory and good) for content differ from task to task. For example, the sequence task requires an explanation of various steps in a recipe while the persuasion task requires building good arguments for making a change.

The following four sections describe the characteristics of each level of each dimension for the description, emergency, sequence and persuasion tasks. The basic definitions of poor, satisfactory and good performance are the same as those presented in the Teacher Observation Ratings Guide. However, the points of the scale have been modified to reflect the number of instances of a particular speaking behavior during a single response to a task. Also, the points of the scale have been further defined to reflect the types of performance that are specific to the four tasks. Pay special attention to the characteristics of good delivery, organization and language and the characteristics of poor, satisfactory and good content, because these characteristics differ somewhat from task to task.

## DESCRIPTION TASK

A. Think about your favorite class or extracurricular activity in your school. Describe to me everything you can about it so that I will know a lot about it. (How about something like a school subject, a club or a sports program?)

### Delivery

POOR

Problems with volume, rate and/or articulation

SATISFACTORY

Absence of problems with delivery and absence of characteristics of good delivery

GOOD

Characteristics of good delivery, typically lively, enthusiastic delivery

### Organization

POOR

Unclear relationships and/or mixed up order

SATISFACTORY

Absence of problems with organization and absence of characteristics of good organization

GOOD

Characteristics of good organization, typically stating favorite subject or activity (My favorite class is ...), using transitions from one topic to another and summarizing aspects of the class or activity

## Content

POOR

Little or irrelevant information about the class or activity and/or poor adaptation to the listener, typically using words or concepts that the listener might not know

SATISFACTORY

Enough information to get a general idea of the class or activity

GOOD

Elaboration about aspects of the class or activity and/or adaptation to the listener, typically explaining terms or using examples

## Language

POOR

Mistakes in grammar or choice of words and/or bland, simplistic language

SATISFACTORY

Absence of problems with language and absence of characteristics of good language

GOOD

Characteristics of good language, typically vivid, descriptive language

## EMERGENCY TASK

B. Imagine that you are home alone and you smell smoke. You call the fire department and I answer your call. Talk to me as if you were talking on the telephone. Tell me everything I would need to know to get help to you. (Talk directly to me; begin by saying hello.)

### Delivery

POOR

Problems with volume, rate and/or articulation

SATISFACTORY

Absence of problems with delivery and absence of characteristics of good delivery

GOOD

Characteristics of good delivery, typically very clear, distinct delivery

### Organization

POOR

Unclear relationships and/or mixed up order

SATISFACTORY

Absence of problems with organization and absence of characteristics of good organization

GOOD

Characteristics of good organization, typically stating the most important information first and presenting the information in a very organized way

## Content

POOR

Little information about the emergency and missing the problem and/or location

SATISFACTORY

A clear statement of the problem (fire, smoke) and location (address)

GOOD

Elaboration about aspects of the emergency (type of fire, type of house or apartment, directions) and/or adaptation to the listener, typically providing information that would be particularly helpful or making sure the listener understands the message

## Language

POOR

Mistakes in grammar or choice of words and/or bland, simplistic language

SATISFACTORY

Absence of problems with language and absence of characteristics of good language

GOOD

Characteristics of good language, typically very precise language

## SEQUENCE TASK

C. Think about something you know how to cook. Explain to me step by step how to make it. (How about something like popcorn, a sandwich or eggs?)

### Delivery

POOR

Problems with volume, rate and/or articulation

SATISFACTORY

Absence of problems with delivery and absence of characteristics of good delivery

GOOD

Characteristics of good delivery, typically lively but clear, distinct delivery

### Organization

POOR

Unclear relationships and/or mixed up order

SATISFACTORY

Absence of problems with organization and absence of characteristics of good organization

GOOD

Characteristics of good organization, typically stating the purpose (I'm going to explain how to make ...) and enumerating the various steps (First ..., Second ...)

## Content

POOR

Little information about the recipe and/or poor adaptation to the listener, typically using words or concepts that the listener might not know

SATISFACTORY

A general explanation of the recipe

GOOD

Elaboration about aspects of the recipe (ingredients, utensils, steps, results) and/or adaptation to the listener, typically explaining terms

## Language

POOR

Mistakes in grammar or choice of words and/or bland, simplistic language

SATISFACTORY

Absence of problems with language and absence of characteristics of good language

GOOD

Characteristics of good language, typically descriptive but precise language

## PERSUASION TASK

D. Think about one change you would like to see made in your school, like a change in rules or procedures. Imagine I am the principal of your school. Try to convince me that the school should make this change. (How about something like a change in the rules about hall passes or the procedures for enrolling in courses?)

### Delivery

POOR

Problems with volume, rate and/or articulation

SATISFACTORY

Absence of problems with delivery and absence of characteristics of good delivery

GOOD

Characteristics of good delivery, typically emphatic delivery

### Organization

POOR

Unclear relationships and/or mixed up order

SATISFACTORY

Absence of problems with organization and absence of characteristics of good organization

GOOD

Characteristics of good organization, typically stating desired change, enumerating reasons and summarizing arguments

## Content

POOR

Few reasons and/or poor adaptation to the listener, typically providing reasons that are speaker centered, not listener centered

SATISFACTORY

Several general reasons for the change

GOOD

Elaboration of reasons, typically argument by example, facts or authority and/or adaptation to the listener, typically to the specific values or motivations of the listener

## Language

POOR

Mistakes in grammar or choice of words and/or bland, simplistic language

SATISFACTORY

Absence of problems with language and absence of characteristics of good language

GOOD

Characteristics of good language, typically emotional or compelling language

# APPENDIX B

## TEACHER BACKGROUND QUESTIONNAIRE

MASSACHUSETTS TEST OF BASIC SKILLS

SPEAKING

Teacher Background Questionnaire

We need you to give us some background information about yourself in order to analyze the results of the speaking ratings. ALL RESPONSES WILL BE KEPT CONFIDENTIAL. Please print the following information and check the boxes which apply to you.

Name: _____

School: _____

---

☐ ☐ ☐ ☐

1. What sex are you?

   ☐ Male

   ☐ Female

2. What subject do you teach primarily?

   ☐ English/Speech

   ☐ Math

   ☐ Science

   ☐ Social Studies/History

   ☐ Other _____

3. Have you ever taken a course or workshop in teaching oral communication skills, e.g., interpersonal communication, group discussion or public speaking?

   ☐ Yes

   ☐ No

Circle one number for each speaking dimension.

| | Poor | Satisfactory | Good |
|---|---|---|---|
| **Practice 1** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Practice 2** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Practice 3** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |

Circle one number for each speaking dimension.

| | Poor | Satisfactory | Good |
|---|---|---|---|
| **Sample 1** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Sample 2** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Sample 3** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Sample 4** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |
| **Sample 5** | | | |
| Delivery (volume, rate & articulation) | 1/2 | 3 | 4/5 |
| Organization (relationships & order) | 1/2 | 3 | 4/5 |
| Content (quantity, relevance & adaptation) | 1/2 | 3 | 4/5 |
| Language (grammar & choice of words) | 1/2 | 3 | 4/5 |

# APPENDIX C

## TEACHER OBSERVATION RATINGS FORM

MASSACHUSETTS TEST OF BASIC SKILLS

SPEAKING

Teacher Observation Ratings

Please print the following information:

Student's Name: _____

Rated By: _____

School: _____

---

☐ [9]☐☐ [1]☐☐

Dimension Ratings

Within each speaking dimension, rate the student along the 5-point scale shown below. Circle one number for each speaking dimension.

| | Poor | | Satis-factory | | Good |
|---|---|---|---|---|---|
| Delivery (volume, rate & articulation) | 1 | 2 | 3 | 4 | 5 |
| Organization (relationships & order) | 1 | 2 | 3 | 4 | 5 |
| Content (quantity, relevance & adaptation) | 1 | 2 | 3 | 4 | 5 |
| Language (grammar & choice of words) | 1 | 2 | 3 | 4 | 5 |

Overall Rating

Rate the student's OVERALL speaking skills on the 8-point scale shown below. Circle one number.

| Lowest | | | | | | | Highest |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

For each of the following questions, check the appropriate box.

1. To what extent do you feel you are familiar with the speaking skills of this student?

   ☐ Very familiar
   ☐ Somewhat familiar
   ☐ Not very familiar

2. How often does this student speak in class?

   ☐ Very often
   ☐ Fairly often
   ☐ Once in a while
   ☐ Hardly ever

3. Are there special circumstances that make you unable to rate this student's speaking skills, e.g., the student recently transferred into class, the student is away from class for remedial or accelerated programs or the student is chronically absent?

   ☐ Yes
   ☐ No

   If yes, explain why: _____

   _____

MASSACHUSETTS TEST OF BASIC SKILLS

SPEAKING

Teacher Observation Ratings

Please print the following information:

Student's Name: _____

Rated By: _____

School: _____

---

☐ 9☐☐ 1☐☐

Dimension Ratings

Within each speaking dimension, rate the student along the 5-point scale shown below. Circle one number for each speaking dimension.

| | Poor | | Satis-factory | | Good |
|---|---|---|---|---|---|
| Delivery (volume, rate & articulation) | 1 | 2 | 3 | 4 | 5 |
| Organization (relationships & order) | 1 | 2 | 3 | 4 | 5 |
| Content (quantity, relevance & adaptation) | 1 | 2 | 3 | 4 | 5 |
| Language (grammar & choice of words) | 1 | 2 | 3 | 4 | 5 |

Overall Rating

Rate the student's OVERALL speaking skills on the 8-point scale shown below. Circle one number.

| Lowest | | | | | | | Highest |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

For each of the following questions, check the appropriate box.

1. To what extent do you feel you are familiar with the speaking skills of this student?

   ☐ Very familiar

   ☐ Somewhat familiar

   ☐ Not very familiar

2. How often does this student speak in class?

   ☐ Very often

   ☐ Fairly often

   ☐ Once in a while

   ☐ Hardly ever

3. Are there special circumstances that make you unable to rate this student's speaking skills, e.g., the student recently transferred into class, the student is away from class for remedial or accelerated programs or the student is chronically absent?

   ☐ Yes

   ☐ No

   If yes, explain why: _____

   _____

# APPENDIX D

## ONE-ON-ONE TEST ADMINISTRATOR'S MANUAL

MASSACHUSETTS DEPARTMENT OF EDUCATION

ASSESSMENT OF BASIC SKILLS

# ONE-ON-ONE TEST ADMINISTRATOR'S MANUAL

## Overview

One-on-one ratings will be administered to all eighth grade students who received the teacher observation ratings. For this approach, a one-on-one rater gives the student four specific speaking tasks. Two one-on-one raters will then independently rate the student's response using a set of speaking rating scales similar to the ones used by teachers for the teacher observation approach. One-on-one raters will take turns giving the speaking tasks to students. Specific steps for implementing the one-on-one ratings are presented in the following sections.

## Preparation

### Identifying Students To Be Tested

Prior to the one-on-one testing session, the one-on-one raters will receive a list of students to be tested, along with their identification numbers.

### Materials

One-on-one raters will receive all of the materials needed for the one-on-one assessment prior to testing. These materials are listed below. Before the first testing session, check to make sure you have everything you need.

Both one-on-one raters will receive a Test Administrator's Notebook. Included in this notebook are:

_____ The One-on-One Test Administrator's Manual

_____ The four speaking tasks

_____ The Observation Ratings Guide

_____ The One-on-One Speaking Ratings Supplement

_____ A sample copy of the One-on-One Speaking Assessment Ratings sheet

_____ A sample copy of the Student Background Questionnaire, "Questions about You"

In addition, the raters will receive one envelope per school which contains:

_____ Student background questionnaires, "Questions about You" (60 copies)

_____ An envelope for collecting student background questionnaires and for returning completed materials

_____ One-on-One Speaking Assessment Ratings sheets (10 copies)

_____ Five blank cassette tapes

For each school, the raters will also receive:

_____ A cassette tape recorder with extension cord or batteries

_____ A supply of No. 2 pencils

Scheduling Testing and Notifying Students and Teachers

The one-on-one assessment will take approximately two days with 30 students being assessed each day. Testing takes approximately five minutes for each student. Arrange an appropriate schedule to test all 60 students.

For example, if the periods in your school are 45 minutes long, the maximum number of students you could schedule per period is nine. However, allow some additional time for introductions, answering questions, interruptions and so forth.

If possible, schedule a follow-up session on a third day for students who do not show up at their assigned sessions. If a student still remains absent, make note of this on the list of participating students. No further follow-up sessions need to be scheduled.

Notify the students about the speaking assessment. A sample notification letter is provided in Appendix A. Teachers should also be notified of the testing dates. This can be done over the PA system or by written announcements, which ever is applicable to your school. Students could be instructed to show their notification letters to their teachers to obtain an excuse from class. The notification procedures are provided as suggestions which may be modified to best suit the circumstances within a particular school.

## Selecting a Testing Location

To conduct the one-on-one ratings, you will need a room with a table or desk and three chairs. The room should be quiet and private. You will also need a waiting area nearby where students may wait for their assessment session and complete their background questionnaires.

## Setting Up the Testing Location

Set two chairs a comfortable distance apart, approximately five feet, on either side or next to a corner of a table or desk. Use one chair for the student and one chair for the rater who administers the tasks. The atmosphere should resemble an interview or conference. It should be relaxed and personal rather than formal. The second rater should be seated near enough to the student to hear the student's responses adequately but not too near as to be disruptive to the student and the rater who administers the tasks.

Place the tape recorder on the desk or table near the test administrator's chair. If you are using a microphone, place it on the desk or table near the student's chair but not too close to the student. Sometimes students tend to handle the microphone if it is too close.

Set the recorder on medium tone and medium volume. Ask the first student you assess to talk a little bit and record his or her voice level. Play it back and make any necessary changes in the volume or position of the microphone or recorder.

Have a copy of the list of participating students on the table along with a copy of all of the speaking tasks. You will also need ten rating forms, five for each rater.

Leave an envelope in the waiting area for collecting student background questionnaires.

Task Assignments

All students will be given the one-on-one speaking tasks in the following order:

A) Descriptive Task

B) Emergency Task

C) Sequence Task

D) Persuasion Task

Become Familiar With Tasks

Before testing, become familiar with the tasks so that you can repeat them accurately and naturally without having to read them.

One-on-One Speaking Assessment Ratings Sheet

Both raters will be marking ratings for the students on the One-on-One Speaking Assessment Ratings sheets. Each sheet is divided into twelve sections, six on each side. Use one section for each student. Students' names and identification numbers should be recorded at the beginning of each student's scheduled test session. The task letters have already been typed on the sheet. Numerical ratings (1-5) for each of the four speaking dimensions should be entered in the appropriate column for each task. If the student does not respond to a task, record a zero for all four dimensions of that task.

## Test Administration

### Explain the Test to the Student

At the beginning of each one-on-one test session, introduce yourself (if necessary) and the second rater. Ask the student's name and check it off from the list of students. Both raters should write the student's name and identification number on their rating sheets.

Tell the student:

> You have been identified to participate in a test of basic speaking skills. This test is part of the State Basic Skills Improvement Program. I will be testing your speaking skills in several situations.
>
> I am going to describe to you four speaking situations. I'll ask you to talk about things you're familiar with such as what you do in your spare time or questions about your school, so you won't have any trouble thinking of things to say.
>
> I will be tape recording your responses. I will identify your responses on the tape recording with an identification number, not with your name.
>
> After you talk, I will mark down a rating of your response on a rating sheet. So will (name other rater). Then I will move on and describe another situation. This will take about 5 minutes to complete.
>
> Your performance on these tasks will not be put in your school records. However, it is important that you do your very best on the speaking tasks, so that we will have an accurate picture of how well students can perform the speaking tasks in the test.

Answer any questions the student might have and try to make him or her feel at ease. Then, move on to explain the specific aspects of the test.

Explain the Tasks to the Student

Tell the student:

> All right, I am ready to start. The first thing I need to do is record your identification number on the tape. Then I will give you a little more information about what I want you to do.

Record the student's identification number on the tape. TURN ON THE TAPE RECORDER and say:

> This is student number (state number).

TURN OFF THE TAPE RECORDER and ask the student:

> Have you ever used a tape recorder before?

If the student says "yes," say:

> Well, then, you know it is very easy.

If the student says "no," say:

> Well, it is very easy. You shouldn't have any problem with it. Just talk to me as you normally would.

Continue:

---

Now let me tell you a little about the speaking tasks. I will describe four speaking situations to you. When you respond to the situations, I want you to talk directly to me.

For some tasks you will have to use your imagination. For instance, you may have to imagine that I am the principal or one of your classmates. In one of the tasks you will have to imagine that you are talking to me on the telephone. Remember to talk to me directly just as you would if you were talking to me on the telephone.

Also, when you are responding to the tasks, imagine that I don't know a lot about what you are talking about. I don't know about your school, your town (or city) or about the things you are interested in. So tell me a lot about what you are talking about so that I will be sure to understand exactly what you are saying.

You will have plenty of time to respond to each speaking task. If you want to, you may take a few moments to think about what you want to say before you start talking.

Tell me as much as you wish about the task. If you talk a very long time, I may interrupt you and ask you to move on to the next situation.

Do you have any questions?

---

Answer the student's questions.

Give the Student the Description Task

When you are sure the student has no more questions, say:

| |
|---|
| All right. Let's begin. Here is the first task. |

Give the student the description task (A). When the student appears ready to speak, TURN ON THE RECORDER and record the student's response.

Provide Suggested Topics

If the student has a hard time thinking of a response, give the student the suggested topics listed in parentheses after the task. Wait 10 seconds. If the student still fails to begin talking after you have given the suggestions, say:

| |
|---|
| All right, let's go to the next task. |

Ask the Student To Talk Directly to You

In some speaking tasks, particularly the task that involves talking on the telephone, students sometimes talk about the task (e.g., First, I would tell you my name ...) instead of doing the task (e.g., Hello, my name is ...). If a student starts to talk about the task, stop the student and ask the student to talk directly to you. Some suggestions for correcting this situation are provided in parentheses after the tasks where this may be a problem.

Look at the Student

While the student is responding, look directly at the student. Try to maintain a pleasant but nonresponsive demeanor. Do not engage in behaviors such as head nodding, laughing, or quizzical looks, which might unwittingly provide a student with positive or negative reinforcement.

Let the Student Respond

Let the student say as much as he or she wants to about each task. When the student appears to be finished, wait three seconds, TURN OFF THE TAPE RECORDER and say:

| |
|---|
| Fine. Now here is the next (or last) task. |

If the student halts but appears to have more to say, wait five seconds. If the student doesn't start talking again, say:

| |
|---|
| That's fine. Let's move on to the next (or last) task. |

If the student talks more than one minute but does not appear to be adding anything that would affect his or her rating, you may interrupt the student and say:

| |
|---|
| That's fine for this task. Let's move on to the next (or last) task. |

## Rate the Student's Response

After the student responds to the task both raters should rate the student's response on each of the four speaking dimensions: delivery, organization, content and language, using the 5-point scales described in the "One-on-One Ratings Supplement." Each rater should make independent ratings without discussion. Base your ratings on the student's entire response. Do not be unduly influenced by the first few seconds of his or her response.

Record your ratings (1, 2, 3, 4 or 5) for each dimension in the appropriate column of the One-on-One Ratings form. If the student does not respond to a task, even after you have given the student suggested topics, record a 0 for each dimension. Make a note in the comments section that the student did not respond to the task.

Indicate any other comments you have about the student's performance in the appropriate section. Comment about any disruptions during the student's response, such as a loud noise, someone coming into the room and so on. Also comment upon any unusual behavior exhibited by the student, such as extreme nervousness, speech problems, difficulty understanding or using English and so on.

## Administer Remaining Tasks

Continue administering the remaining three tasks to the student. Administer the emergency task (B) second, the sequence task (C) third and the persuasion task (D) last. Follow the same procedures for all four tasks.

## Changing Tapes and Recording Identification Numbers

Watch the tape recorder to make sure you are not running out of tape. You will reach the end of one side when the counter reads approximately 480. When you have finished one side, mark the identification numbers of the students contained on the tape on the label and turn the tape over.

## Dismiss the Student and Hand out Background Questionnaire

After the student has completed all of the tasks, thank him or her for cooperating and ask the student not to discuss the tasks with other students as this would give the other students an unfair advantage.

Give the student one of the background questionnaires to complete in the waiting area. Explain to the student that all responses will be kept confidential and that the information will be used to help explain the test results. Pencils may be handed out to the students at this time or kept in the waiting area. Ask students to leave their completed questionnaires in the envelope provided in the waiting area.

## Put One-on-One Speaking Assessment Materials into Envelope

When you have completed all of the one-on-one speaking assessments, put the ratings sheets and the student background questionnaires into the return envelope. Matthew Towle will pick up all of the materials following the testing dates.

## APPENDIX A

### Sample Student Notification Letter

Student's Name: _____

You have been identified to participate in a speaking test for the Basic Skills Improvement Program. Not all students in your grade have been asked to participate in this test. You have been selected at random because we need more information about how well this test works, not because we think your speaking skills are poor.

The speaking test will require approximately five minutes of your time. You will be asked to speak only about things that are familiar to you, so you won't have any problem thinking of things to talk about.

The testing will occur on (date) and your scheduled time is (time). Please report to (room) promptly at this time.

If you need to be excused from a class or study hall, show this letter to your teacher and, if necessary, obtain a hall pass from your teacher.

Sincerely,

## APPENDIX E

## ONE-ON-ONE SPEAKING ASSESSMENT RATINGS FORM

Rater: _____

School _____

# ONE-ON-ONE SPEAKING ASSESSMENT RATINGS

| Student's Name | Identification Number | Task | Delivery | Organization | Content | Language | Comments |
|---|---|---|---|---|---|---|---|
| | 1 | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |
| | 2 | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |
| | 3 | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |
| | 4 | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |
| | 5 | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |
| | | A | | | | | |
| | " | B | | | | | |
| | " | C | | | | | |
| | " | D | | | | | |

# APPENDIX F

## STUDENT BACKGROUND QUESTIONNAIRE

MASSACHUSETTS TEST OF BASIC SKILLS

SPEAKING

Questions About You

Please print the following information.

Name: _____ ☐ ☐ ☐

School: _____ ☐

For each of the following questions, check the box which applies to you.

1. What sex are you?

   ☐ Male
   ☐ Female

2. What is your racial/ethnic background?

   ☐ Black
   ☐ Hispanic
   ☐ White
   ☐ Other

3. What language did you FIRST learn to speak at home?

   ☐ English only
   ☐ English and another language
   ☐ Another language

4. Did you transfer into this school since Christmas vacation of this year?

   ☐ Yes
   ☐ No

5. How often have you been given work in school aimed at helping you improve your speaking skills?

   ☐ Hardly ever
   ☐ Once in a while
   ☐ Fairly often
   ☐ Very often

6. How often do you have a chance to speak in your classes?

   ☐ Hardly ever
   ☐ Once in a while
   ☐ Fairly often
   ☐ Very often

Following is a set of statements numbered 7 through 12. Read each statement carefully and decide how much you agree or disagree. Check the appropriate box.

| | | Strongly Agree | Agree | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| 7. | I have no problem "jumping into a discussion" in a group. | ☐ | ☐ | ☐ | ☐ |
| 8. | I have difficulty asking my teachers to clarify or explain things, even when I don't understand them. | ☐ | ☐ | ☐ | ☐ |
| 9. | I am nervous while participating in group discussions. | ☐ | ☐ | ☐ | ☐ |
| 10. | I have no problem talking with people who are much older than I am. | ☐ | ☐ | ☐ | ☐ |
| 11. | I don't like having to call up someone I don't know on the telephone. | ☐ | ☐ | ☐ | ☐ |
| 12. | I find it hard to get a word in when I am in a group discussion. | ☐ | ☐ | ☐ | ☐ |

Project conducted and report written by:

Nancy Mead
Lynn Grover Gisi
Education Commission of the States
Denver, CO

Richard Hill
RMC Research Corporation
Hampton, NH